DOCUMENT RESUME

ED 471 224 TM 034 657

AUTHOR	Bunch, Michael B.
TITLE	Item Review 101: Where We've Been, Where We're Going, How We'll Get There.
PUB DATE	2002-06-00
NOTE	67p.; Paper presented at the Annual Meeting of the Council of Chief State School Officers (Palm Desert, CA, June 2002).
PUB TYPE	Guides - Classroom - Learner (051) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC03 Plus Postage.
DESCRIPTORS	*Large Scale Assessment; *Review (Reexamination); Test Construction; *Test Items; *Testing Programs
IDENTIFIERS	Item Review Scale

ABSTRACT

This module explains test review as it is now performed in most large-scale testing programs. It addresses the fundamental aspects of item review principally for fairness, sensitivity, and bias, and to a lesser degree for content and construct validity. The module is designed for a survey course and thus has a broad, rather than deep, focus. The module is designed to majors and nonmajors alike and is designed to be useful for state and local department of education staff as well as officers and employees of testing organizations. The module contains these sections: (1) definition and purpose; (2) basic issues; (3) historical overview; (4) current practices; (5) case study; and (6) recommendations. A self-assessment for the user is included. Appendixes contain a code of fair testing practices and a discussion of fairness and sensitivity training. (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 471 224

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

M. B. Bunch

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Item Review 101: Where We've Been, Where We're Going, How We'll Get There

Michael B. Bunch
Measurement Incorporated

TM034657

Paper presented at the annual meeting of the Council of Chief State School Officers, Palm Desert, California, June 25, 2002.

BEST COPY AVAILABLE

This module orients the student to test item review as it is now performed in most large-scale testing programs. It addresses fundamental aspects of item review principally for fairness, sensitivity and bias, and, to a lesser degree, for content and construct validity. As this is a survey course, the focus of this module will be broad but not very deep. This module is open to majors and non-majors alike, to state and local department of education staff, as well as officers and employees of testing organizations.

Module Outline

- Definition and purpose
- Basic issues
- Historical overview
- Current practices
- Case study
- Recommendations

A self-assessment and two appendices are included.

Required Reading

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (1999). *Standards for Educational and Psychological Testing*. Washington, DC: AERA. [Especially Part II, Chapters 7 - 10]

Joint Committee on Testing Practices (1988) *Code of Fair Testing Practices in Education*. Washington DC: American Psychological Association. (See Appendix A)

Supplemental Reading

Berk, R. A. (1982). Handbook of Methods for Detecting Test Bias. Baltimore, MD: Johns Hopkins University Press.

Cole, N. S. & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational Measurement* (*3rd Ed.*). Washington, DC: National Council on Measurement in Education.

Definition and Purpose

At some point in the development of any large-scale assessment, and typically at multiple points, individuals or groups review the test items to make sure they are fair, free

of bias, and devoid of content that would prove upsetting or appear to be insensitive. The terms *fairness*, *sensitivity*, and *bias* are sometimes used interchangeably. While the terms are similar, there are both philosophical and practical differences in their meanings. Cole (1981), for example, noted that many researchers tend to use the term *fairness* to refer to a broad range of social policy issues related to tests and test use and the term *bias* to refer to the more technical (validity) issue that could be subjected to statistical analysis. This conceptualization survives, at least officially.

Fairness applies to a broad range of social policy issues having to do with overall appropriateness of tests and test items: readability, reading load, vocabulary, grade-level appropriateness, age-level appropriateness, freedom from bias, avoidance of sensitive and sensitized issues, and a host of mechanical issues (print size and font, page layout, and format familiarity, for example). A test can be completely unbiased and even fairly sensitive to individual and group differences and still be unfair simply by being uniformly unfair to all students. A test must therefore satisfy more criteria to be considered fair than it would have to satisfy to be considered unbiased or insensitive. In its most fundamental form, fairness is an issue of basic test construction, of content and construct validity (cf. Payne, 1997).

Sensitivity is a relatively recent addition to the fairness review criterion set and relates almost solely to test content that one or more groups of individuals may find offensive. It also addresses subtle and not-so-subtle underlying assumptions of the test developers. The Educational Testing Service (ETS, undated), in its *Fairness Review* offers an example of underlying assumptions that deals with an item from a social worker certification exam that seems to imply that social workers' clients are members of minority groups, but that social workers themselves are not. Such an item would not be sensitive to the feelings of minority candidates. In practice, sensitivity encompasses a rather broad range of issues, some of which may not relate to groups. For example, while death is a natural part of life and a recurring theme in literature, many developers of literature tests avoid reference to death because some test takers may have recently lost loved ones. To introduce this theme or even the word into a high-stakes test might prove so upsetting to an adolescent test taker, that the results for that individual would be invalid. Other themes, phrases, and words produce similar results and are avoided for similar reasons. Unlike *fairness*, which addresses broad social policy issues, *sensitivity* tends to be somewhat more individual and therefore more difficult to define.

Bias is perhaps the most thoroughly researched of the three terms. *Educational Measurement* (*3rd Edition*) devotes an entire chapter to this topic (Cole & Moss, 1989), and countless articles and books have been written on the subject. Bias has been reduced to a variable that can be examined statistically (cf. Berk, 1982; Camilli & Shepard, 1994). Given the scope of the bias literature, it is not surprising that a clear and simple definition of the term is elusive. Shepard (1982), for example, offered the following conclusion to a chapter tantalizingly entitled "Definitions of Bias":

> In keeping with Scriven's earlier advice regarding the loss of meaning in operational definitions, a simplistic, concrete definition of item bias has not been offered. Instead, an understanding of bias-as psychometric features

> that somehow misrepresent the abilities of one group-is expected to guide the detection of bias in particular instances. (p. 25)

Cole & Moss (1989) take a strictly technical approach to bias within the larger context of test validity and offer the following definition under the heading "The Technical Meaning of Bias":

> Bias is present when a test score has meanings or implications for a relevant, definable subgroup of test takers that are different from the meanings or implications for the remainder of the test takers. Thus, *bias is differential validity of a given interpretation of a test score for any definable, relevant subgroup of test takers* (p. 205, italics in the original).

This definition of bias is reflected in the *Standards for Educational and Psychological Testing* (see **Required Reading**, above):

> Bias ... is said to arise when deficiencies in a test itself or the manner in which it is used result in different meanings for scores earned by members of different identifiable subgroups. (p. 74)

Two important concepts emerge from these two definitions:

1. Bias is viewed strictly in technical terms and not necessarily in the traditional or dictionary terms that suggest prejudice or malice. In this regard, bias is not the same thing as lack of fairness.

2. Bias is restricted to identifiable subgroups of examinees, rather than individuals, a notion that carries enormous implications.

We will not address issues related to bias in test administration or interpretation in this module. Instead, we focus only on bias that is inherent in the test items or other test content (e.g., reading passages, graphics, directions). For the purpose of this module, our operational definition of bias is as follows:

Bias is any feature of a test that creates a systematic disadvantage for any identifiable subgroup of examinees. Disadvantage is defined as any noticeable difference in outcome that is not directly related to the construct being tested.

Further delineation of the elements of this definition may be helpful:

Systematic disadvantage - defined strictly in terms of test scores or likelihood of a correct response. Suppose two groups known to be identical in mathematical ability respond to an algebra item. Fifty percent of Group A answer the item correctly, while 70 percent of group B answer the item correctly. Further, answer choice D (which is incorrect) contains a vocabulary term that is not relevant to the algebra problem but which is known to discriminate between the two groups (i.e., it is a term well known to members of group B). Suppose, finally, that far more members of group A than group B select D as the

correct response. The presence of this irrelevant vocabulary term creates a systematic disadvantage for members of group A because it disproportionately attracts members of this group in an irrelevant way.

Identifiable subgroup - quite possibly the crux of the matter. Historically, members of racial and ethnic minorities have satisfied this description. Females have also qualified. Students with handicapping conditions qualify under certain conditions. In technical terms, these groups have been referred to as "target" groups, while white, male examinees have been referred to as "reference" groups. The inclusion of this phrase in the definition raises some interesting and not entirely academic questions: Are all left-handed students an identifiable subgroup? Are members of fundamentalist religious groups an identifiable subgroup for the purpose of statewide science tests? How large or encompassing does a group have to be to qualify for this special status? How many identifiable subgroups can be considered at one time?

Noticeable difference - In the algebra example, the difference between the two groups was 20 percentage points, fairly noticeable in anyone's book. What if the difference had been two or three points? Fortunately, we have statistical techniques for answering such questions.

Not directly related to the construct being measured - Returning to our algebra example, if the difference between the two groups had been due to a difference in mathematical or algebraic ability, there would have been no bias, because the difference is related to the construct being measured. What made the item biased (or at least appear to be biased) was the fact that an irrelevant feature of the item created the score discrepancy. This concept is extremely important when analyzing group differences on item and test performance. Simple (even significant) differences between groups do not automatically mean that the test or item is biased. One viable explanation that must be considered is the possibility that the two groups actually differ with respect to the construct being tested.

No definition of test item bias would be complete without a discussion of differential item functioning (DIF). If a whole test yields different results for different groups of examinees, as, for example, a college admissions test underpredicting the likelihood of success of minority applicants or overpredicting the likelihood of success for majority applicants, the test is said to demonstrate differential validity. That is, the power of the test to predict college success differs, depending on the group of applicants to which it is administered. In similar fashion, an item that yields different results for different subgroups of examinees, even though those two groups are similar or equivalent with respect to the construct of interest, is said to demonstrate DIF. For example, suppose males and females of comparable verbal ability respond to a reading test item. Further suppose that 80 percent of the female examinees answer the item correctly, and 60 percent of the male students answer the item correctly. At first glance, one might conclude that the item demonstrates DIF. Indeed, if the samples are large enough so that the 80 percent and 60 percent estimates are stable, the item may be demonstrating DIF. Specific statistical techniques (discussed later in this module) have been developed to detect DIF.

Admittedly, this is a rather lengthy definition (or perhaps explanation) of the term *bias*. It is important to be thorough, however, because this term tends to attract the greatest amount of attention and is the source of more extensive debate than the other two terms. As will be noted later in this module, there are statistical methods associated with bias, while there are none associated with either fairness or sensitivity. Since many people equate statistical technique with scientific rigor, it is important to know what all the fuss is about.

Basic Issues

Reviewing tests and test items for fairness, sensitivity, and bias involves four basic issues:

- Why does this need to be done?
- Who will do it?
- What rules will they follow?
- What will happen to their findings?

It is the responsibility of the oversight agency (i.e., the state or local department of education) to answer these questions. While answers sometimes evolve over time, it is preferable to answer these questions before any test items are developed. It may be necessary to answer some or all of these questions repeatedly, and occasionally with emphasis. These questions are addressed below in fairly broad terms. Later sections of this module take up some of these questions in more detail.

Why does this need to be done? Review for fairness, sensitivity, and bias is part of an overall quality control process that is designed to ensure the construct validity of the test. Fairness issues are addressed when content experts and individuals familiar with student populations assert that the test measures only the constructs of interest in a manner that is consistent with the developmental maturity of the students taking the test; i.e., that the test possesses construct validity. Furthermore, failure to conduct such a review of test items invites criticisms that the test is inherently biased or worse, even if no flaws exist in the test. In short, it is good public relations to ask a group of qualified individuals to review test items for any feature that would systematically create a disadvantage for any group of students. However, even if no public relations value were to accrue, the value of such a review, strictly in terms of test validity, would be tremendous.

Who will do it? Typically, state and local departments of education (and test publishers) employ groups of educators with content expertise and several years of experience teaching students in the target and reference groups. These groups often have the status of a standing committee. Others are *ad hoc*. In a few instances, content experts (mostly classroom teachers and district curriculum supervisors) who review the content of the test items also evaluate the items in terms of fairness, sensitivity, and bias.

In some instances panels of individuals are selected for their leadership roles in various identifiable subgroups. Michigan, for example, has a Bias Review Committee that includes educators, parents, and community members. In Ohio, the Fairness/Sensitivity Review Committee has few teachers; most members are representatives of organizations such as the Alliance of Black School Educators, National Organization for Women, and various student advocacy groups.

The choice of who will review the items and tests is the responsibility of the testing agency. This is a matter requiring careful consideration of even more questions:

- How large should the group be?
- Should reviewers constitute a standing committee, or should there be a different group for each review?
- If there are standing committees, how long should members serve?
- Should membership terms be staggered in order to preserve continuity?
- Should members be educators, non-educators, or a mix?
- What should be the minimum educational and experience qualifications for group membership?

What rules will they follow? Groups and committees seem to work best when there is a set of rules to follow. All too often, agency staff leave review committees to their own devices to develop their own rules. Occasionally, no one establishes rules, and chaos reigns. Establishing the charter of the review committee and ground rules for the conduct of reviews is the responsibility of the testing agency. The charter identifies the scope and nature of the committee's task and authority. For example, if there is a standing content review committee which deals with grade-level appropriateness, grammatical accuracy, and overall relevance to content standards, it is probably redundant for fairness/sensitivity/bias committees to address these features of test items. A clear line of demarcation between the two committees is helpful.

The authority of the group would also be clearly spelled out at this time. Will the review committee or the testing agency have final authority to keep or reject an item? This is not a trivial issue.

Having defined the scope and nature of the work they are to perform, the agency should then identify procedures for item review. Such procedures include, but are not limited to, the following:

- Group vs. individual review
- Design and completion of review forms
- Initial and ongoing training
- Item outcome options (e.g., reject, modify, accept, other)
- Dealing with disagreements
- Reviewing items without statistics
- Reviewing items with statistics
- Reviewing reading passages, graphics, and other materials
- Protocol

The first issue is quite fundamental. Some testing agencies assemble groups of educators to review items and permit discussion among the reviewers, but the task of reviewing the items is essentially an individual one; i.e., each member completes an item review form (and may or may not sign it) and turns it in. In other instances, members debate each item, and a chair signs a final copy of the accepted, modified, or rejected item. Thus, the design of review forms is quite central to the overall process of reviewing, because the form captures the full scope of the task (you can't rate it if it's not on the

form) as well as the flow (e.g., individual signatures vs. signature of the chair).

Training is important because it establishes or reinforces the charter of the group as well as operational procedure. To attempt to conduct content review without initial and ongoing training is ill advised. To attempt to conduct fairness/sensitivity/bias review without initial and ongoing training is to invite disaster. All other issues in the list above are set out in training and reinforced by review meeting facilitators. To the extent that all procedures are spelled out (with written copies distributed to all members), review sessions can proceed fairly smoothly.

What will happen to their findings? This question was addressed tangentially in the context of spelling out the group's authority. Will the findings of the reviewers simply be part of the documentation of test development, or will the reviewers have veto power over each and every item. Or will they operate in some middle ground? If the testing agency is to have final authority with regard to the retention or rejection of test items, that fact needs to be clearly communicated to the group early and often. These facts can be fairly simply communicated: The group recommends; the agency acts on the recommendations. Sometimes, this means modifying or rejecting a group recommendation. The rationale for this position is that the agency has to stand behind the test because its name is on the cover, and it bears sole responsibility for the validity and legal defensibility of the test.

If the group has only advisory status, what becomes of its findings? This is a matter that the testing agency should consider very carefully. If the agency is seen to be cavalier with regard to committee recommendations, the quality of those recommendations, and indeed of the process, will deteriorate. The wise course seems almost always to be to select members carefully, allow a fairly broad course, accept nearly all recommendations, and choose carefully those fights that must be won. When it is necessary to overrule a committee's recommendation, the agency must have a solid argument that does not violate either the charter or the procedures established for the group. If it becomes clear that rules need to be changed, it is best to change them after a recommendation has been overruled rather than during the process.

Historical Overview

In the beginning, there was testing. And it was good. Then came group intelligence testing. Then came test items requiring children to identify as pretty a woman who looked distinctly western European, even if some of those children were of eastern or non-European ancestry. Then darkness covered the face of the deep, and it was not so good.

A few years ago, I became aware that many in our industry lack a sense of history of assessment. I first noticed it in conversations with staff of other testing companies, specifically with regard to the aforementioned test that discriminated against children with mothers who looked non-western European. Then I noticed it in my own staff, people I had hired and trained. After thinking about this situation for a while, I came to the

following conclusion: We are hired to do specific work. If that work has any historical context, that context can usually be measured in months, such as a testing program authorized by a piece of legislation passed last year. That's as far back as anyone has the time or inclination to look. No one pays us to think about the past or the future, only the present. For my purposes here, next spring's test is more present than future. When I refer to the future, I refer to something many years out.

James McKeen Cattell at Columbia and Lewis Terman at Stanford pioneered scientific principals of psychological and educational testing in the early 1900s. Almost as soon as they began publishing their views and developing tests, the anti-testing movement arose. The *Atlantic Monthly* carried a running debate between Lewis Terman and Walter Lipman in the late 1920s and 1930s. Lipman, the well-known author and social critic, opposed the use of tests because of their association with eugenics.

The debate over testing continued in the pages of the *Atlantic Monthly* and other highbrow publications into the 1940s. The War temporarily shifted people's attention to more immediate concerns, but the attack on testing resumed in the 1950s, this time in the hands of Banesh Hoffman, who published *The Tyranny of Testing* in 1962. His basic criticism of testing was almost identical to the current one expressed by Alfie Kohn (2000) in *The Case Against Standardized Testing*: the tests don't test anything important and are actually harmful to education because they force teachers to focus on unimportant things at the expense of meaningful instruction. The educational community's overdependence on multiple-choice test items was at the heart of the matter.

In 1976, Nicholas Leman published *The Big Test: The Secret History of the American Meritocracy*, a lengthy account of ETS and its founders, Henry Chauncy and James Conant, former professor and president of Harvard, respectively. Leman described Chauncy as a man in love with the notion of the efficiency of the multiple-choice test and the efficient classification of candidates to the nation's most prestigious institutions of higher learning. While Leman credited Chauncy with attempting to broaden the outreach and admissions practices of schools such as Harvard and Princeton, his primary focus was on the lock-step, fill-in-the-bubble approach that characterized ETS and its products. Leman attacked not so much a specific test or test item but an approach to creating and implementing testing programs.

Right on the heels of the Leman critique of Chauncy the man and ETS the organization, Ralph Nader joined the fray. The Nader organization sponsored a report by Allan Nairn entitled "The Reign of ETS: The Corporation that Makes Up Minds," published in 1980. Nairn argued that the SAT did not predict college freshman grades and was actually a better predictor of family income. The tone of the report was essentially "The SAT: Unsafe At Any Speed."

Last, but certainly not least on this list, is the eminent physicist/philosopher Stephen Jay Gould, who published *The Mismeasure of Man* in 1981. *The Mismeasure of Man* was a bestseller and recipient of the coveted Critics Circle Award for general nonfiction in 1981. Gould took test developers and users to task for constructing hypotheses of human intellect and then manipulating results statistically to make sure their

presuppositions were supported. The attack focused on the science of psychometrics as well as on the racist intentions of the psychometricians.

For most of the 20th century, through attacks and criticisms, the mechanics of test development remained a mystery to most people. Security was the watchword of every testing agency and a way of life for every testing agency employee and test user. Even the critics had to rely on partial knowledge of the tests they criticized. Lipman, Hoffman, Nairn, Gould, and Kohn criticized tests and testing practices from an outsider's point of view. The concept of item validity, while psychometrically established, had yet to embed itself in the consciousness of the anti-testers.

In the midst of the attack on tests by intellectuals came an attack in the courtrooms. The *Larry P. v. Riles* decision in 1979 was directly concerned with the labeling of disproportionate numbers of minority children in California as mentally handicapped or learning disabled. The larger impact was that it finally established as legal fact what Lipman, Hoffman, and others had argued for decades. More importantly, it opened the doors and windows of the testing industry and allowed sunlight to touch the vaunted intelligence test. The whole world was able to see how such tests were made and used. Other legal challenges of this period (1970s and 1980s) similarly addressed the use of tests in employment (*Griggs v. Duke Power*), college admission (*Regents of the State of California* v. *Bakke*), and *Debra P. v. Turlington* (disproportionate rate of failure of minority students on Florida's Statewide Student Assessment Test).

At about the same time that individuals and organizations were pursuing judicial relief from the testing industry, state lawmakers were weighing in. In July 1979, New York passed the first so-called "truth-in-testing" law. Taking effect in 1980, the law required all publishers and administrators of tests within the state of New York to reveal the entire contents of the tests after examinees had received their scores. Individual examinees could thus review each and every test item and its correct answer. While the law was more or less aimed at ETS and specifically at the SAT, it has had much wider impact, including the state's own Regents Exam and other statewide tests.

The educational testing community was set back by the *Larry P.* decision, but not overly concerned about the future because we were beginning to rely less on IQ tests and more on educational achievement tests. Similarly, this community did not become overly concerned by other judicial decisions of this era: *Griggs v. Duke Power* (employment testing; doesn't concern us), or *Bakke* (college admissions; doesn't concern us), or *Diana v. California State Board of Education* (ESL; we have a much broader focus). *Debra P. v. Turlington* and New York's truth-in-testing law brought it all home.

The *Debra P.* decision focused squarely on one of the newly minted competency tests sweeping the country in the 1970s. Truth in testing meant that anyone could examine any test item on any test. Concerned parents and anti-testers could take more time to pick apart items than all the editors and content reviewers ever had. State departments of education and directors of assessment branches had to take notice because they all had a stake. Suddenly all large-scale, high-stakes tests were laid bare for detailed public inspection. Since the public cannot generally be counted upon to have a high

degree of psychometric sophistication, they focused on the things they did understand: the content of individual test items. Much of what they saw appalled them. Much of what appalled them should have. Much more should not have.

Test item review as we now know it was thus born amid controversy, animosity, residual guilt, and anxiety about what to do next. As is so often the case, many in the educational community overreacted. Test item review committees were established, often without regard to mission or method, and handed the keys to the store. As a result, some committees made matters worse. Admirably, others recognized the opportunity to serve the community and made wise, informed decisions about tests and test items, often in a way that even the most astute psychometricians could not.

During the past 20 years, review of test items for fairness, sensitivity, and bias issues has become common practice for large-scale assessments. The 1985 edition of the *Standards* (AERA/APA/NCME, 1985), contained specific guidelines for detecting and reducing bias at the item level. As those *Standards* were being drafted and reviewed, the American Psychological Association (APA) formed two working committees: the Committee on Professional Standards and the Committee on Psychological Standards and Assessments (Diamond & Fremer, 1989). These two committees were keenly aware of the criticisms recently leveled against testing and the court cases recently adjudicated or under review.

These two committees sought much broader input and support and brought together representatives of several test publishers as well as representatives of the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME). This meeting gave rise to the Joint Committee on Testing Practices (JCTP). One of the two principal documents produced by the JCTP was the *Code of Fair Testing Practices in Education* (JCTP, 1988). The Code spells out roles and responsibilities for both test developers and test users. State departments of education and other educational agencies are often included in both categories because they contract with testing companies to develop tests and use results of those tests to make decisions about individual students, schools, and districts. These roles and responsibilities are summarized in Table 1.

Table 1
Roles and Responsibilities of Test Developers and Test Users
(from *Code of Fair Testing Practices in Education*)

Test Developers	Test Users
Define and explain the test's content	Demonstrate appropriateness of the test
Provide understandable score reports	Inform users about students' rights
Read and understand descriptions of the test content	Avoid inappropriate use of the test
Inform students of their rights	

In 1999, the three national organizations (AERA, APA, and NCME) issued a new edition of the *Standards* (AERA, APA, NCME, 1999). All of Part II of that document (Fairness in Testing) is devoted to fairness in testing and test use (Chapter 7), the rights and responsibilities of test takers (Chapter 8), testing individuals with diverse linguistic backgrounds (Chapter 9), and testing individuals with disabilities (Chapter 10), a total of 38 pages. Thus, as criticisms of tests and test developers have mounted over the years, so too have guidance and useful suggestions. Indeed, the introduction to Chapter 7 of the 1999 *Standards* provides an extremely useful summary and review of the fundamental concepts before laying out 12 standards. Chapters 8-10 are similarly constructed, with excellent background pieces preceding a total of 36 clearly delineated standards. While the majority of this section is devoted to fairness of test use at a gross level, some of the standards relate specifically to test items:

- **7.3** - When credible research reports that differential item functioning exists across age, gender, racial/ethnic, cultural, disability, and/or linguistic groups in the population of test takers in the content domain measured by the test, test developers should conduct appropriate studies when feasible. Such research should seek to detect and eliminate aspects of test design, content, and format that might bias test scores for particular groups.

- **7.4** - Test developers should strive to identify and eliminate language, symbols, words, phrases, and content that are generally regarded as offensive by members of racial, ethnic, gender, or other groups, except when judged to be necessary for adequate representation of the domain.

- **7.10** - When the use of a test results in outcomes that affect the life chances or educational opportunities of examinees, evidence of mean test score differences between relevant subgroups of examinees should, where feasible, be examined for subgroups for which credible research reports mean differences for similar tests. Where mean differences are found, an investigation should be undertaken to determine that such differences are not attributable to a source of construct underrepresentation or construct-irrelevant variance. While initially the responsibility of the test developer, the test user bears responsibility for uses with groups other than those specified by the developer.

While analysis of items is clearly specified in Standards 7.3 and 7.4, it may take a bit of digging to find such a specification in Standard 7.10. Let us assume that group mean differences have been found. This standard calls for further investigation of sources of those differences. After looking to the groups for answers, the test developer is left with the items as the only other source of information.

Current Practices

Where there are large-scale testing programs, there are test item reviews. With the ready availability of the 1999 *Standards* and the *Code of Fair Testing Practices in*

Education, as well as a 20-year history of item review, one would expect a fair amount of uniformity in what is reviewed and how it is reviewed. As seems to be the case in all other state-by-state comparisons, differences abound. A quick scan of state department of education Web sites for the terms *bias*, *fairness*, and *sensitivity* shows a variety of procedures and approaches to dealing with test item review. The following summary relies on such scans to some degree but also includes more in-depth descriptions of the practices of the states which Measurement Incorporated (MI) currently serves.

Arkansas - The Arkansas Comprehensive Testing, Assessment, and Accountability Program (ACTAAP) encompasses a statewide accountability system with tests at grades 4, 6, 8, and high school. Very soon, grades 3, 5, and 7 will be added to the program. For all tests in this program, separate content committees review items for each subject at each grade level. The Arkansas Bias Review Committee (BRC) is composed of 20 school administrators selected by the Arkansas Department of Education. Committee members represent a cross section of the student population, including those who have special needs. The BRC examines materials (items and passages) to identify perspectives considered sensitive or biased by the general public and as a result should be avoided or carefully addressed on the ACTAAP assessments. These perspectives include, but are not limited to, ethnic, gender, religious, cultural, regional and socioeconomic biases.

California - The Public School Accountability Act of 1999 established the Academic Performance Index. This act also codified the policy of test fairness.

Connecticut - Separate content and bias/sensitivity review committees examine all items developed for the various statewide testing programs. The bias/sensitivity review committee is composed of 15 members, primarily classroom teachers. Their responsibility is to review all potential items and make recommendations to the Connecticut State Department of Education regarding retention, modification, or rejection of items.

Florida - The Florida Comprehensive Assessment Test (FCAT) is reviewed by several committees, including content, bias, and sensitivity. Bias and sensitivity are separate committees with separate charges and compositions. While the bias committee is made up primarily of educators, the sensitivity committee includes parents, community members, and other non-educators.

Georgia - The Student Assessment Handbook describes the test development process in some detail, including the test item review process. For the Georgia High School Graduation Tests and the Criterion Referenced Competency Tests, committees of Georgia educators meet to review items simultaneously for content and bias issues; i.e., one committee performs both reviews for a given grade and subject.

Illinois - The Technical Report for the Illinois Statewide Achievement Test (ISAT) describes in some detail the test development process, including the test item review process.

Maryland - A Maryland Statewide Proficiency Assessment Program (MSPAP) Question & Answer document (dated March 2001) addresses test development and review

procedures.

Michigan - The Michigan Educational Assessment Program (MEAP) includes several different test item review committees. The Bias Review Committee (BRC) is made up of Michigan teachers, parents, and community leaders. The Michigan Department of Education has established 35 criteria by which the BRC will judge items. Overriding concerns are fairness and offensiveness. Specific categories of potentially offensive material include evocation of unpleasant emotion, mention of illegal or immoral behavior (e.g., gambling), invasions of privacy, references to politics, reinforcement of prejudices, items with direct or indirect religious references (e.g., Halloween, the Bible, etc.), and social class issues.

New York - The State Education Department (SED) has published its own Standards (September 2001) and Sensitivity Guidelines. These standards and guidelines apply to the New York Regents Exam, the Regents Exam Component Retest, and other statewide exams. Reviews for bias and sensitivity are typically conducted outside the test development cycle. In many instances, one contractor develops the items, and another is responsible for assembling the committees and conducting validity studies. Results of the studies are then forwarded to the SED which, in turn, forwards them to the development contractor.

Ohio - The Ohio Department of Education (ODE) Web site contains a downloadable copy of the *Code of Fair Testing Practices in Education*. The ODE mission statement includes a very clear reference to test fairness as part of that mission. The newly authorized Ohio Graduation Test (OGT) is being developed under new guidelines which call for a fairness/sensitivity review panel with much broader membership and scope than the one that had been responsible for the development of the previous test. See the **Case Study** for further details.

Pennsylvania - The Department's Web site contains a downloadable document entitled "Principles, Guidelines, and Procedures for Developing Fair Assessments" (September 1999). This 87-page manual describes in considerable detail the test development process and the roles and responsibilities of review committees. It is a model of a state-developed guide to test development.

Wyoming - The duties of the Bias & Sensitivity Committee are spelled out in an Equity Review document. This committee, like the ones in Michigan and Ohio, includes both educators and community leaders.

In a typical contract, MI staff interact with item review committees three times during a single test development cycle: passage review, new item review, and field test review. Each stage involves different objectives and strategies.

Passage review. As copyright permissions have become more and more expensive, we have instituted a separate review stage for reading and language arts items. Before any items are written and before any royalties are paid to authors or publishers, MI staff present several candidate passages. The review committees pore over these passages

and identify those which may be offensive in whole or in part. Those passages deemed unworthy, for whatever reason, are eliminated from the pool. MI staff then make arrangements to pay for copyright permission for accepted passages and assign passages to item writers.

Occasionally, a passage is deemed mostly appropriate, but a single word or phrase is found to be offensive. The committee sometimes requests that we ask the author or publisher for permission to alter the text. Sometimes the copyright holder consents, sometimes not. In one celebrated case, the author happened to be Ray Bradbury, well-known science fiction author and outspoken free-speech advocate. The passage in question contained the word "damn." The committee agreed to approve the passage if the offending word could be excised. We contacted the publisher, who referred us to Mr. Bradbury. We made the request in a carefully worded letter. Mr. Bradbury's response was quite terse: "Hell, no!"

Recent newspaper accounts suggest that similar requests have come from New York's sensitivity review committees, but that the requests have been honored without the involvement of authors and publishers. Jeanne Heifetz, statewide co-chair of the Parents Coalition to End High Stakes Testing and the mother of a New York high school senior, discovered that several reading passages in the Regents Exam had words missing. A closer inspection of several recently released tests showed that most passages had words missing. Comparison to the originals revealed that references to religion, alcohol, nudity, race, sex, and even mild profanity had been excised. Subsequent investigations revealed that these changes had been made without author or publisher permission. After a series of articles and editorials in the *New York Times* and the *New York Post*, Education Commissioner Richard P. Mills reversed the SED's position on 'sanitizing' reading passages. In fact, the SED had apparently already begun to back away from the practice, in that the June 18-19, 2002, edition of the Regents Exam contained unexpurgated texts.

New item review. MI typically presents new items to committees several months before scheduled field testing. Items are presented one per page on review forms. For the fairness/sensitivity/bias (FSB) review committees, these review forms contain check-off boxes or lines for racial/ethnic, gender, religious, age, SES, cultural, and other bias. Committees review and discuss items and point out any offensive or stereotypical language. At the end of the discussion of each item, the committee takes a vote to accept, modify, or reject the item. The committee chair logs the vote and signs the form.

Frequently, offensive items can be salvaged by making minor modifications. Occasionally, items cannot be saved. In many instances, the FSB committee reviews all items before content committees examine them. Any items rejected by the FSB committee are removed from the set of items to be reviewed by content committees, and any modifications suggested by the FSB committee are made before the content committees see them.

In some instances (e.g., Georgia), a single group reviews for content and bias. In Georgia, there are no committees as such; each reviewer completes his or her own form and signs it. If there are 12-15 reviewers, it is possible that only 8-10 would review a

given item. We group items in packets, either by reading passage (for Language Arts) or by content standard (Mathematics, Science, and Social Studies). We provide an orientation to the review task by explaining each of the criteria shown in Table 2 and then describe the process step by step. Each reviewer takes a packet and a rating form, reviews all the items in the packet, completes the rating form, and returns the packet to the head table. During the process, reviewers are free to discuss any rating with any other reviewer, and there are frequent large-group discussions. At the end of the two-day review session, each item has been reviewed by 8-10 different reviewers or by the whole group if an item is particularly interesting or troubling. At the conclusion of the process, we aggregate the ratings and forward the results to the Georgia Department of Education, along with our recommendations.

Table 2
Georgia High School Graduation Test Item Selection Criteria

- **Match specification?**
- **Covered?**
- **Clear?**
- **Correct?**
- **Distractors Plausible?**
- **Unbiased?**
- **Passage OK?**
- **Graphic OK?**

Post field-test review. After analysis of field test data, we conduct a final round of item reviews with content and bias committees. At this time, we present difficulty and discrimination data to the content review committees and DIF data to bias review committees. Table 3 shows a typical data set for a bias review committee.

Table 3
Sample Field Test Data for Bias Review Committee

Item	Std	Pos		Sex	Race	Key	P/Mean	Corr	A	B	C	D	
	Omit												
16	013	1			R	B	0.81	0.38	2	81	5	9	
	4												
06	082	2				B	0.52	0.36	3	52	10	32	
	4												
11	082	3			R	D	0.81	0.33	10	5	2	81	
	3												
02	053	4				D	0.78	0.38	8	7	3	78	
	4												
18	063	5				B	0.33	0.19	7	33	6	49	
	4												
22	035	6				A	0.71	0.40	71	12	10	3	
	4												

23	082 4	7			R	C	0.71	0.43	13	8	71	4	
08	035 3	8				B	0.64	0.49	5	64	21	6	
28	084 5	9				C	0.42	0.17	6	34	42	14	

Q#	P-White		P-Black		W-A	W-B	W-C	W-D	WRESP		B-A	B-B	
	B-C	B-D	BRESP										
1	0.88	0.65	1	88	3	6	428	3	65	8	18	173	
2	0.61	0.34	2	61	7	29	427	4	34	16	38	172	
3	0.87	0.65	8	3	1	87	429	16	8	5	65	174	
4	0.86	0.64	6	5	2	86	427	13	12	5	64	174	
5	0.38	0.25	6	38	5	49	426	10	25	10	48	172	
6	0.80	0.54	80	10	7	1	426	54	15	18	9	176	
7	0.81	0.51	11	4	81	3	426	21	16	51	8	175	
8	0.75	0.41	5	75	16	3	428	5	41	36	112	175	
9	0.45	0.34	5	38	45	9	423	8	27	34	25	172	

Q#	P-Male	P-Female	M-A	M-B	M-C	M-D	MRESP	F-A	F-B	F-C	F-D		
	FRESP												
1	0.82	0.80	3	82	6	8	373	1	80	5	11	393	
2	0.53	0.50	2	53	9	34	372	4	50	11	31	391	
3	0.82	0.80	10	3	3	82	372	9	6	2	80	394	
4	0.81	0.76	7	6	4	81	371	9	8	3	76	392	
5	0.32	0.35	7	32	7	50	368	8	35	5	49	391	
6	0.74	0.68	74	13	7	3	370	68	11	13	4	392	
7	0.75	0.67	12	7	75	3	367	15	9	67	5	391	
8	0.69	0.61	5	69	18	6	370	6	61	23	7	394	
9	0.45	0.39	5	34	45	12	367	6	35	39	15	389	

Legend

Item - the item's bank position (not test position)
Std - the content standard addressed by the item
Pos - position in the test booklet (item number)
Sex - flag for significant Mantel-Haenszel statistic by sex
Race - flag for significant Mantel-Haenszel statistic for race (Black/White)
Key - Correct answer (CR for constructed-response item, A-D for multiple-choice)
P/Mean - overall p value or raw score mean for the item
A-D - percentage of students choosing each response option
Omit - percentage of students omitting this item
Q# - same as Pos (item number)
P-White - percentage of white students answering correctly
P-Black - percentage of black students answering correctly
W-A - W-D - percentage of white students selecting options A-D
WRESP - total number of white students responding to this item
B-A - B-D - percentage of black students selecting options A-D
BRESP - total number of black students responding to this item
P-Male - percentage of male students answering correctly
P-Female - percentage of female students answering correctly
M-A-M-D - percentage of male students selecting options A-D
MRESP - total number of male students responding to this item
F-A-F-D - percentage of female students selecting options A-D
FRESP - total number of female students responding to this item

The flags for sex and race (black/white only for this sample) are based on Mantel-Haenszel statistics. MI uses the three categories described by Camilli & Shepard (1994), which they attribute to Michael Zieky (1993):

A - no DIF
B - possible DIF
C - probable DIF

We start with the flagged items and provide the by-group statistics as backup. We also provide the base test statistics by group for comparison. For example, if a committee is surprised to see a 10-point difference in p values for black and white students, their concern can be somewhat moderated by the fact that the base test shows an 8-point difference, and the base test contains only those items that this committee has previously deemed to be bias-free.

In a typical post-field-test review, committees examine the items and data, concentrating more on the items than the data. Typically, they will remember certain items and the discussions they had during new item review and will check the statistics to see if they were right. In the end, few items are rejected by the bias committees at post-field-test review.

Case Study: Ohio Graduation Test[1]

Ohio's Senate Bill 1 (SB1), signed into law in 2001, authorized the Ohio Graduation Test (OGT) as well as other changes in the state's testing programs. The bill mandated a well-orchestrated program of public engagement and authorized the establishment of a Fairness/Sensitivity Committee to be a partner in the test development process. Section 3301.079(G) of the Ohio Revised Code spells out the role and responsibility of the Committee: *fairness sensitivity review committee must not allow any question on any achievement or diagnostic test, or any proficiency test, to include, be written to promote, or inquire as to individual moral or social values or beliefs. The decision of the committee shall be final. (line 368)*

Other roles of the Fairness/Sensitivity Committee, as defined by ODE staff include the following:

- ensure that items do not disadvantage groups of students because of their race, ethnicity, gender, or disability.
- ensure that diverse cultures are represented in assessments and that material used is neither offensive nor stereotypical of any student group.

To ensure that no one group of students is unfairly advantaged or disadvantaged some examples of questions committee members should reflect on are listed below:

- Do items reflect an improper balance between racial or ethnic groups and not maintain cultural integrity of those groups?
- Is there an uneven balance between sexes - numerically and in terms of the significance and prominence of the activity?
- Do illustrations reflect an imbalance of physical types and avoid evidence of physical disability?
- Do items reflect materials understood only by specific cultural groups?
- Will the language used in the item be interpreted differently by members of different groups?
- Would items be found offensive or emotionally disturbing to a group of students?
- Do item directions or scoring guidelines assist or credit responses more typical of one group of students than another?
- Do scoring guidelines reserve the highest score for those students who provide more information than actually requested, than a less test-wise student?

The above questions are given only as a sample to help determine the existence of bias. Ohio's diverse population must be considered in reviewing all potential items with sensitivity to all aspects of bias (i.e., race, gender, ethnicity, religion, disability, language, socio-economic status)

This committee may also be required to advise the development of other assessment support materials.

The Committee is made up of Ohio citizens, some of whom are educators, and some of whom represent specific constituencies. These include

Dayton Christian Schools, Inc
Ohio NOW (State affiliate of the National Organization of Women)
Kiwanis International
Alliance of Black School Educators
Diocese of Columbus Catholic Schools
Islamic Academy

In addition to the representatives from these organizations, the Committee includes four teachers, three parents, two professors, two local Board of Education representatives, and one ESL coordinator.

Committee ground rules have been clearly spelled out in a document entitled "Ohio Fairness/Sensitivity Committee Membership Guidelines." These are listed below:

- One person may talk at a time so that each committee member's concern(s) can be heard and considered.
- Share questions and concerns openly.
- Remain respectful and sensitive to the views of other committee members.
- Honor decisions made by the committee.
- Use meeting time to complete committee tasks only (use break or lunch time to discuss other educational issues).
- Individual contributions are not to be shared outside the meeting.

The "Guidelines" also spell out the test development process, length of service, meeting notification and reimbursement, and ethical issues and provide a glossary of assessment terms.

The Ohio Fairness/Sensitivity Committee had its organizational meeting on May 7, 2002. ODE staff presented the charge and overall organization of the Committee. The author then led a training session designed to help members understand their roles and responsibilities and the nature of large-scale, high-stakes test development and use. The presentation is included as Appendix B. What had originally been planned as a 90-minute presentation turned into a day-long activity with numerous questions and many lengthy discussions. By the end of the day, it was clear that the Committee was committed to its task and keenly interested in every detail of test construction.

The Committee had its first opportunity to review reading passages and test items on June 10-12, 2002. The first day was devoted to reviewing reading passages carried forward from the previous testing program. All passages had been previously approved by content and bias review committees. Nevertheless, the Fairness/Sensitivity Committee rejected three reading passages (and all associated test items). Their reasoning proved compelling. One passage was rejected because it was about baseball. More specifically, the items required knowledge about baseball not contained in the passage. The previous committee had approved the passage and items, even though some general knowledge of baseball was assumed by many of the items. One F/S Committee member (the ESL

coordinator) pointed out the large numbers of immigrant children who have no knowledge of American sports at all. The other members agreed and urged MI to continue to include sports-related passages but to look for greater diversity in the types of sports portrayed. Another passage was rejected because of age bias (an elderly woman, not Miss Daisy, hires a driver).

In general, the Committee held to its appointed task. At one point during the review of mathematics items, one Committee member took out a calculator and began to work the problem. The MI facilitator reminded the member that content was the purview of the content review committee, and he quickly remembered that he did not need to work the problem. Quick glances from other members reinforced his recollection. Concerns about content did not simply disappear, however. At various points, Committee members raised content issues. The MI facilitator took each concern, made arrangements to share them with the content review committees who met the week of June 17-21, 2002, and to provide a report to the F/S Committee at its next session. This arrangement satisfied the Committee.

While it may be too early to tell exactly how the Ohio Fairness/Sensitivity Committee will fare, indications to date are positive. The roles and responsibilities have been clearly spelled out in a written document shared with all Committee members; the members have had a thorough orientation to their tasks, and they have demonstrated a willingness and ability to carry out these tasks in a professional manner.

Recommendations

In the 20 or so years that have passed since the inception of the current version of test item review, there have been many improvements to the original model(s). Many state testing programs are characterized by well-documented charters, policies, and procedures governing the actions of test item review committees. Pennsylvania and Ohio stand out as examples of this documentation. In many cases, the responsibilities of item review committees are commensurate with their authority.

In other cases, however, policies and procedures do not exist, or if they do, state staff and committee members seem to be unaware of them. Contractors, who work *for* the state agencies and *with* the committees, often do not know the histories of the committees and therefore make incorrect assumptions about the work they do and how they do it.

With the passage of *No Child Left Behind* and the imminent development of hundreds of new tests, we find ourselves at a crossroads. Before moving ahead, it may be helpful to take a long and careful look back, note where we are and how we got here, carry the best of the past 20 years into the future and make concrete plans to avoid the mistakes of the last two decades as we establish review processes for the next generation of tests. The following recommendations are based on twenty years of working with content and fairness/sensitivity/bias review committees in a dozen states and close observation of practices that work and some that don't.

1. Define the roles and responsibilities of reviewers, and adhere to them firmly. As noted above, chaos can ensue if review committees do not know their roles, responsibilities, or limits of authority. The Ohio case study illustrates some of the ways to avoid chaos. Major elements of those roles and responsibilities include the following:

Scope - Fairness/sensitivity/bias review committees typically review materials only for fairness/bias/sensitivity issues. They typically do not or should not take over the role of content review committees. If they do have content concerns, there should be a mechanism for forwarding those concerns to the content review committee and for reporting back to the FSB committees the actions of the content committees.

Authority - Typically, all committees serve in an advisory capacity. Legally constituted authorities such as boards of education and their paid staff (i.e., state department of education staff) have the responsibility to develop the tests. Authority should be commensurate with responsibility; those responsible should have the authority to carry out their responsibilities. We have sometimes found ourselves in the situation in which state department of education staff lived in fear of review committees who seemed to have absolute authority. Once the chair signed the form, no one, including the state superintendent, could alter a syllable. Such a situation is clearly indicative of a dysfunctional system.

Tenure - Some committees seem to have lifetime memberships. Such memberships may be by design or by default. At the other extreme are those committees with no continuity at all from one review session to the next. The testing agency should clearly define a term limit, as well as qualifications for serving on the committee.

2. Set process and procedure rules, and adhere to them. Having established the roles and responsibilities of reviewers, it would be quite helpful to establish a set of ground rules by which they will carry out their responsibilities. Most of these ground rules can be taken care of through the careful design and introduction of review forms. Such forms should include all the allowable discussion topics (i.e., types of bias), with the clear implication that anything not on the form is out of bounds. Think carefully before adding 'Other' to the list. Having created forms, the final step is how to complete the form and how to move through a set of items. Will each reviewer complete and sign a form, or will the group vote and permit a chair to sign a single form for all? Should there be discussion after each item or after a set of related items? How long can discussion of one item last? How can a committee member or staff member halt debate and move on to the next item? How will the committee handle rewrites or revisions? Should such items be revised and brought back for another review?

These and similar questions must be addressed. Having answered each question and created a set of ground rules, it is best to commit these rules to writing and to share them with all committee members. It is then a good idea to remind committee members of the rules at the start of each review session. As new members come onto the committees, it may be useful to devote a significant amount of review time to training and retraining. Finally, it is important to enforce the rules in a consistent manner. A rule violated twice

without consequence is no longer a rule.

3. Define terms. The attached training presentation defines *bias* in both positive and negative terms. Members of the Ohio Graduation Test Fairness/Sensitivity Committee were instructed before they saw the first OGT item that some things (delineated) were sources of bias, and other things (also delineated) were not. The panel had an opportunity to discuss why some things (e.g., gender stereotyping) led to item bias and why other things (e.g., things that bug me personally) do not. By the end of the first training session, all members of the committee had a clear understanding not only of what did and what did not constitute bias but also, by transfer, of their roles and responsibilities.

4. Ground all policies and procedures in reality. There are currently nearly six billion versions of reality. It may be possible to narrow those down to a handful and then to one for the purposes of one fairness/sensitivity/bias review committee. The *Standards* and *Code of Fair Testing Practices in Education* are a good starting place. An understanding (or at least an appreciation) of copyright law is also helpful. The current demand for authentic tests and authentic reading passages means that we need to put passages from real, published literature into tests. Real literature has generally not been scrubbed to meet the demands of many sensitivity committees. At some point, given the recent experiences of the New York SED, a meaningful debate on what can and should be done to literary passages targeted for inclusion in large-scale assessments is in order. Other states may be in a similar situation right now, waiting for someone to discover their sanitized reading passages.

Beyond the issue of passage fidelity, it is necessary to consider how realistic it is to allow anyone at any time to declare any topic or word objectionable. In Ohio and elsewhere, we have focused on **identifiable groups**. One specific example we used in Ohio had to do with timber wolves. The fact that timber wolves attacked someone's grandmother some time in the distant past is not sufficient reason to reject a passage about timber wolves.

5. Train and retrain committee members. Everyone knows what bias is. Who needs to be trained to recognize it? The problem is that everyone *has* a bias, and *no one* really knows what it is. Few people walk in off the street knowing what the AERA/APA/NCME *Standards* are or what the *Code of Fair Testing Practices in Education* is, and yet these documents are central to the work of fairness/sensitivity/bias review committees. Moreover, while we would not dream of turning even experienced test item writers loose with assignments for a new testing program without extensive training, we seem perfectly content to let review committee members fend for themselves, make their own ever-changing rules, and make uninformed decisions which they erroneously believe have the force of law. Every new committee should be thoroughly trained before the first item or passage review. Each succeeding item or passage review session should begin with retraining, ranging from an hour to half a day, depending on the tasks to be undertaken. Each member should be required to bring his or her guidelines to each meeting, and the meeting facilitators should have extra copies for those who forget.

Training is particularly important in the final stage of item review when item

statistics are shared. Members can refer to the Hansel-Gretel statistic all they want, as long as they know that C means that they should take a very close look and that five percent of the time, C happens for no apparent reason. Group mean differences on items denote different things in different contexts. Differential attractiveness of specific item distractors is important. Gaining an understanding of these concepts is very important. For non-statistically oriented groups who have this experience but once a year, it is very unlikely that retention of the statistical concepts will be high. Retrain; review, and reexamine.

6. Be prepared for attacks. Much of what we do today in terms of test and item review can be traced directly to criticisms of tests in the last century. Many of the criticisms were justified, and we as a profession have made great strides toward addressing them. Many other criticisms were unjustified or exaggerated. We need to know the difference. That requires that we spend some time educating ourselves and our colleagues on the following topics:

- The history of testing, including the various anti-testing movements
- The NRT/CRT evolution at the item, test, and statistical analysis levels
- Legal issues related to testing
- Statistical techniques for detecting test and item bias
- AERA/APA/NCME *Standards*
- *The Code of Fair Testing Practices in Education*
- Applicable state law

Challenges will come from many quarters. Test item review committee members have legitimate questions and concerns that someone in a position of authority (preferably an agency staff member, but a contractor will do in a pinch) should be able to answer clearly and satisfactorily. When these questions are ignored or glossed over, the committee member assumes the worst. When criticisms from parents, the local media, and mass-market media arise, our potential allies are not available because we have not won them over to our side. We need to be prepared to provide positive leadership to committee members, to solicit their support with their colleagues back home, and to make our case as often as possible (with their help) to the public at large that the tests are fair and free of bias because decent, hard-working people just like them are advising the test developers.

Inasmuch as many of the criticisms of testing over the past century have been unfounded or exaggerated, and certainly persistent (even those that have been thoroughly disproved), we have an opportunity each time we meet with item review committees to dispel anew the more entrenched false notions. Without subjecting committee members to long, boring history lessons or attempting to browbeat them into psychometric orthodoxy, we could address each of the following topics in an informal and friendly way and then prove that we are committed to the concepts we are promoting:

- We really are trying to help students, honest!
- It is not our intention to create or perpetuate racial, gender, or social class inequality.
- We are not even particularly interested in maximizing differences among students.

- The multiple-choice item is not the work of the devil.
- Authenticity is not limited to constructed-response items and essay tests.
- Tests are designed to sample from a universe of generalizable student behaviors.
- There are no secret caves or hideouts where we write the real tests.
- We are not trying to tell children (or teachers) what to think, or even how to think.
- We are not assessing moral values or beliefs.
- If you find a mistake, consider what your life would be like if every word you uttered were reported on the front page of the *New York Times* or the *Boston Globe*, and act accordingly.

If we demonstrate to review committees that we are listening to them, they might just listen to us. How do we do this? The second time we meet with a review committee, we put in front of them materials that clearly show that the recommendations they made at the first meeting were incorporated not only into the items they reviewed at that meeting but have been generalized to the new set of items. We then continue to follow this pattern at subsequent meetings. Along the way, we address the issues listed above, as teachable moments arise.

[1] Thanks to Tom Bulgrin of the Ohio Department of Education for the "Ohio Fairness/Sensitivity Committee Membership Guidelines" and assistance in preparing this section.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). *Standards for Educational and Psychological Testing*. Washington, DC: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: Author.

Berk, R. A. (1982). *Handbook for Detecting Test Bias*. Baltimore, MD: Johns Hopkins University Press.

Camilli, G. & Shepard, L. A. (1994). *Methods for Identifying Biased Test Items*. Thousand Oaks, CA: SAGE Publications.

Cole, N. S. (1981). Bias in testing. *American Psychologist, 36* (10), 1067 - 1077.

Cole, N. S. & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.) *Educational Measurement (3rd Edition)*. Washington, DC: American Council on Education.

Debra P. v. Turlington. 564 F. Supp. 177 (M. D. Fla. 1983).

Diamond, E. E. & Fremer, J. (1989). The Joint Committee on Testing Practices and the Code of Fair Testing Practices in Education. *Educational Measurement: Issues and Practices, 8* (1), 23-24.

Educational Testing Service (undated). *Overview: ETS Fairness Review*. Princeton, NJ: Author.

Gould, S. J. (1981). *The Mismeasure of Man*. New York: Norton.

Griggs v. Duke Power Co. 401 U. S. 424 (1971).

Hoffman, B. (1962). *The Tyranny of Testing*. New York: Crowell-Collier Press.

Joint Council on Testing Practices (1988). *Code of Fair Testing Practices in Education*. Washington, DC: National Council on Measurement in Education.

Kohn, A. (2000). *The Case Against Standardized Testing: Raising the Scores, Ruining the Schools*. Portsmouth, NH: Heinemann.

Larry P. v. Riles. 495 F. Supp. 926 (M. D. Cal. 1979); appeal docketed, No. 80-4027 (9th Cir. Jan. 17, 1980).

Leman, N. (1976). *The Big Test: The Secret History of the American Meritocracy*. New York: Farrar, Straus, and Giroux).

Nairn, A. & Associates (1980). *The Reign of ETS: The Corporation That Makes Up Minds*. Washington, DC: Ralph Nader Institute.

PASE v. Hannon. 506 F. Supp. 831 (N. D. Ill. 1980).

Payne, D. A. (1997). *Applied Educational Assessment*. Belmont, CA: Wadsworth Publishing Company.

Regents of the University of California v. Bakke, 438 U.S. 265 (1978).

Shepard, L. A. (1982). Definitions of bias. In R. A. Berk (Ed.) *Handbook for Detecting Test Bias*. Baltimore, MD: Johns Hopkins University Press.

Zieky, M. (1993). Practical questions in the use of DIF statistics in item development. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning: Theory and Practice* (pp. 337 - 364). Hillside, NJ: Lawrence Erlbaum.

Self-Assessment

1. Which of the following would qualify as an identifiable subgroup?

A. female students
B. left-handed yak herders
C. individuals with criminal records
D. people who feel 'sort of queasy' when they take tests

2. Which statement best describes the relationship between bias and validity?

A. A biased item reduces the validity of a test.
B. A test with a biased item cannot be valid.
C. Validity is impervious to bias in individual items.
D. Bias affects predictive, but not construct, validity.

3. Who is responsible for eliminating bias from tests?

A. bias review committees
B. state testing agencies
C. contractors
D. all the above

4. How are rules for the conduct of item review committees created?

A. Committees make them up as they go along.
B. Testing agencies make them up as they go along.
C. Committees make them up at the first meeting and then stick to them.
D. Testing agencies make them up before the first meeting and then stick to them.

5. What bearing does the Otis-Lennon Intelligence Scale have on current test item review practices?

A. The publisher of the Otis-Lennon pioneered most of today's item review practices.
B. Criticisms of the Otis-Lennon opened the door to criticism of all tests and their individual items.
C. Current item review practices are designed to align statewide testing programs with the Otis-Lennon and similar tests.
D. Otis and Lennon taught many of the individuals who now manage test development companies.

6. Which statement best describes the impact of *Debra P. v. Turlington* on modern statewide assessment programs?

A. It was a landmark Supreme Court decision that specified how and when states could test.
B. The plaintiffs in that case directly implicated testing programs in several other states.

C. The final ruling effectively barred any further lawsuits against statewide testing programs.
D. The methods used by Florida Department of Education staff to defend the test have been adopted by other state departments of education.

7. According the *Code of Fair Testing Practices in Education*, which role does a state testing agency play?

A. test developer only
B. test user only
C. both test developer and test user
D. neither test developer nor test user

8. What is indicated by a Mantel-Haenszel Category C classification of a test item?

A. More than the expected number of minority students chose C as the correct answer.
B. The item clearly demonstrates differential item functioning.
C. The item does not discriminate.
D. The item is definitely biased.

9. Which state recently made headlines for altering reading passages in tests to eliminate offensive language?

A. California
B. Florida
C. New York
D. Wyoming

10. What is the most likely reason that one fairness/sensitivity review committee would identify an unfair passage about baseball, while another would not?

A. One committee was made primarily of women, while the other was made up primarily of men.
B. One committee included a teacher who specializes in non-native speakers of English, while the other did not.
C. One committee focused only on bias issues, while the other focused on bias and fairness.
D. One committee reviewed several sports-related passages, while the other only reviewed a few sports-related passages.

Key: 1 - A; 2 - A; 3 - D; 4 - D; 5 - B; 6 - D; 7 - C; 8 - B; 9 - C; 10 - B

Appendix A

Code of Fair Testing Practices in Education

CODE OF FAIR TESTING PRACTICES IN EDUCATION
PREPARED BY THE JOINT COMMITTEE ON TESTING PRACTICES

The Code of Fair Testing Practices in Education states the major obligations to test takers of professionals who develop or use educational tests. The Code is meant to apply broadly to the use of tests in education (admissions, educational assessment, educational diagnosis, and student placement). The Code is not designed to cover employment testing, licensure or certification testing, or other types of testing. Although the Code has relevance to many types of educational tests, it is directed primarily at professionally developed tests such as those sold by commercial test publishers or used in formally administered testing programs. The Code is not intended to cover tests made by individual teachers for use in their own classrooms.

The Code addresses the roles of test developers and test users separately. Test users are people who select tests, commission test development services, or make decisions on the basis of test scores. Test developers are people who actually construct tests as well as those who set policies for particular testing programs. The roles may, of course, overlap as when a state education agency commissions test development services, sets policies that control the test development process, and makes decisions on the basis of the test scores.

The Code has been developed by the Joint Committee on Testing Practices, a cooperative effort of several professional organizations, that has as its aim the advancement, in the public interest, of the quality of testing practices. The Joint Committee was initiated by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. In addition to these three groups the American Association for Counseling and Development/Association for Measurement and Evaluation in Counseling and Development, and the American Speech-Language-Hearing Association are now also sponsors of the Joint Committee.

This is not copyrighted material. Reproduction and dissemination are encouraged. Please cite this document as follows:

> Code of Fair Testing Practices in Education. (1988)
> Washington, D.C.: Joint Committee on Testing Practices.
>
> (Mailing Address: Joint Committee on Testing Practices,
> American Psychological Association, 1200 17th Street, NW,
> Washington, D.C. 20036.)

The Code presents standards for educational test developers and users in four areas:

A. Developing/Selecting Tests
B. Interpreting Scores
C. Striving for Fairness
D. Informing Test Takers

Organizations, institutions, and individual professionals who endorse the Code commit themselves to safeguarding the rights of test takers by following the principles listed. The Code is intended to be consistent with the relevant parts of the *Standards for Educational and*

Psychological Testing (AERA, APA, NCME, 1985).[1] However, the Code differs from the Standards in both audience and purpose. The Code is meant to be understood by the general public; it is limited to educational tests; and the primary focus is on those issues that affect the proper use of tests. The Code is not meant to add new principles over and above those in the Standards or to change the meaning of the Standards. The goal is rather to represent the spirit of a selected portion of the Standards in a way that is meaningful to test takers and/or their parents or guardians. It is the hope of the Joint Committee that the Code will also be judged to be consistent with existing codes of conduct and standards of other professional groups who use educational tests.

A. DEVELOPING/SELECTING APPROPRIATE TESTS*

Test developers should provide the information that test users need to select appropriate tests.

TEST DEVELOPERS SHOULD:

1. Define what each test measures and what the test should be used for. Describe the population(s) for which the test is appropriate.

2. Accurately represent the characteristics, usefulness, and limitations of tests for their intended purposes.

3. Explain relevant measurement concepts as necessary for clarity at the level of detail that is appropriate for the intended audience(s).

4. Describe the process of test development. Explain how the content and skills to be tested were selected.

5. Provide evidence that the test meets its intended purpose(s).

6. Provide either representative samples or complete copies of test questions, directions, answer sheets, manuals, and score reports to qualified users.

Test users should select tests that meet the purpose for which they are to be used and that are appropriate for the intended test taking populations.

TEST USERS SHOULD:

1. First define the purpose for testing and the population to be tested. Then, select a test for that purpose and that population based on a thorough review of the available information.

2. Investigate potentially useful sources of information, in addition to test scores, to corroborate the information provided by tests.

3. Read the materials provided by test developers and avoid using tests for which unclear or incomplete information is
provided.

4. Become familiar with how and when the test was developed and developed and tried out.

5. Read independent evaluations of a test and of possible alternative measures. Look for evidence required to support the claims of test developers.

6. Examine specimen sets, disclosed tests or samples of questions, directions, answer sheets, manuals, and score reports before selecting a test.

TEST DEVELOPERS SHOULD:

7. Indicate the nature of the evidence obtained concerning the appropriateness of each test for groups of different racial, ethnic, or linguistic backgrounds who are likely to be tested.

8. Identify and publish any specialized skills needed to administer each test and to interpret scores correctly.

TEST USERS SHOULD:

7. Ascertain whether the test content and norm group(s) or comparison group(s) are appropriate for the intended test takers.

8. Select and use only those tests for which the skills needed to administer the test and interpret scores correctly are available.

*Many of the statements in the Code refer to the selection of existing tests. However, in customized testing programs test developers are engaged to construct new tests. In those situations, the test development process should be designed to help ensure that the completed tests will be in compliance with the Code.

B. INTERPRETING SCORES

Test developers should help users interpret scores correctly.

TEST DEVELOPERS SHOULD:

9. Provide timely and easily understood score reports that describe test performance clearly and accurately. Also, explain the meaning and limitations of reported scores.

10. Describe the population(s) represented by any norms or comparison group(s), the dates the data were gathered, and the process used to select the samples of test takers.

11. Warn users to avoid specific, reasonably anticipated misuses of test scores.

12. Provide information that will help users follow reasonable procedures for setting passing scores when it is appropriate to use such scores with the test.

13. Provide information that will help users gather evidence to show that the test is meeting its intended purpose (s).

Test users should interpret scores correctly.

TEST USERS SHOULD:

9. Obtain information about the scale used for reporting scores, the characteristics of any norms or comparison group(s), and the limitations of the scores.

10. Interpret scores taking into account any major differences between the norms or comparison groups and the actual test takers. Also take into account any differences in test administration practices or familiarity with the specific questions in the test.

11. Avoid using tests for purposes not specifically recommended by the test developer unless evidence is obtained to support the intended use.

12. Explain how any passing scores were set and gather evidence to support the appropriateness of the scores.

13. Obtain evidence to help show that the test is meeting its intended purpose (s).

C. STRIVING FOR FAIRNESS

Test developers should strive to make tests that are as fair as possible for test takers of different races, gender, ethnic backgrounds, or different handicapping conditions.

TEST DEVELOPERS SHOULD:

14. Review and revise test questions and related materials to avoid potentially insensitive content or language.

15. Investigate the performance of test takers of different races, gender, test and ethnic backgrounds when samples of sufficient size are available. Enact procedures that help to ensure that differences in performance are related primarily to the skills under assessment rather than to irrelevant factors.

16. When feasible, make appropriately modified forms of tests or administration procedures available for test takers with handicapping conditions. Warn test users of potential problems in using standard norms with modified tests or administration procedures that result in non-comparable scores.

Test users should select tests that have been developed in ways that attempt to make them as fair as possible for test takers of different races, gender, ethnic backgrounds, or handicapping conditions.

TEST USERS SHOULD:

14. Evaluate the procedures used by test developers to avoid potentially insensitive content or language.

15. Review the performance of test takers of different races, gender, and ethnic backgrounds when samples of sufficient size are available. Evaluate the extent to which performance differences may have been caused of the test.

16. When necessary and feasible, use appropriately modified forms or administration procedures for test takers with handicapping conditions.
Interpret standard norms with care in the light of the modifications that were made.D.
INFORMING TEST TAKERS

Under some circumstances, test developers have direct communication with test takers. Under other circumstances, test users communicate directly with test takers. Whichever group communicates directly with test takers should provide the information described below.

TEST DEVELOPERS OR TEST USERS SHOULD:

17. When a test is optional, provide test takers or their parents/guardians with information to help them judge whether the test should be taken, or if an available alternative to the test should be used.

18. Provide test takers the information they need to be familiar with the coverage of the test, the types of question formats, the directions, and appropriate test-taking strategies. Strive to make such information equally available to all test takers. Under some circumstances, test developers have direct control of tests and test scores. Under other circumstances, test users have such control. Whichever group has direct control of tests and test scores should take the steps described below.

TEST DEVELOPERS OR TEST USERS SHOULD:

19. Provide test takers or their parents/guardians with information about rights test takers may have to obtain copies of tests and completed answer sheets, retake tests, have tests rescored, or cancel scores.

20. Tell test takers or their parents/guardians how long scores will be kept on file and indicate to whom and under what circumstances test scores will or will not be released.

21. Describe the procedures that test takers or their parents/guardians may use to register complaints and have problems resolved.

Note: The membership of the Working Group that developed the Code of Fair Testing Practices in Education and of the Joint Committee on Testing Practices that guided the Working Group was as follows:

Theodore P. Bartell	John J. Fremer	George F. Madaus	Nicholas A. Vacc
John R. Bergan	(Co-chair, JCTP	(Co-chair, JCTP)	Michael J. Zieky
Esther E. Diamond	and Chair, Code	Kevin L. Moreland	
Richard P. Duran	Working Group)	Jo-Ellen V. Perez	(Debra Boltas and
Lorraine D. Eyde	Edmund W. Gordon	Robert J. Solomon	Wayne Camara of
Raymond D. Fowler	Jo-Ida C. Hansen	John T. Stewart	the American
	James B. Lingwall	Carol Kehr Tittle	Psychological
		(Co-chair, JCTP)	Association
			as staff

served

liaisons)

Additional copies of the Code may be obtained from the National Council on Measurement in Education, 1230 Seventeenth Street, NW, Washington, D.C. 20036. The Web site is http://www.apa.org/science/jctpweb.html. Single copies are free.

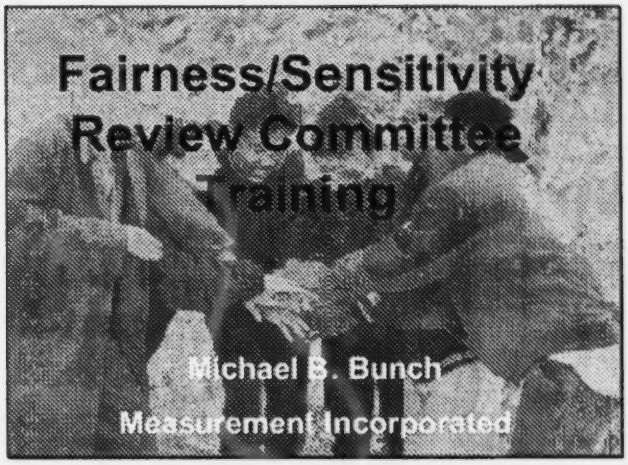
The Joint Committee on Testing Practices, in conjunction with several other organizations, has also produced a videotape entitled ***"The ABC's of School Testing"*** designed to help parents understand the many uses of testing in schools today. Various types of tests and their appropriate uses in the school setting are illustrated, and aptitude and achievement tests are also discussed. In addition to the videotape, two publications are also included: *Leader's Guide* and the *Code of Fair Testing Practices in Education*. To obtain an order form for this video package, contact the Science Directorate at 202-336-6000 or at testing@apa.org, or contact the National Council on Measurement in Education (NCME) at 202-223-9318.

[1] These Standards have been supplanted by the 1999 publication of *Standards for Educational and Psychological Testing.*

Appendix B

Fairness/Sensitivity Review Committee Training

(Powerpoint Notes Pages)


Fairness/Sensitivity
Review Committee
Training
Michael B. Bunch
Measurement Incorporated

What is item bias?

Item bias is any feature of a test item that takes the focus off the content of the item and places it on an irrelevant difference between groups of test takers.

The key is that **bias is a source of invalidity** of test items and tests. To the extent that a reading test measures something other than reading (e.g., gender or race), that reading test is invalid. For example, a reading test that contained only passages about sports and cars might be more interesting to boys than to girls. If boys and girls of otherwise equal reading ability happened to get different results on this "reading" test, we could probably conclude that the test was less a measure of reading ability and more a test of interest in sports and cars. To make the test less biased and more valid, we would balance the passages to present topics that interested both boys and girls or some that interested one group and some that interested the other.

What is NOT item bias?

- **Poor item construction**
- **Content that an individual finds offensive**
- **Differential access to instruction**
- **Mean group differences**

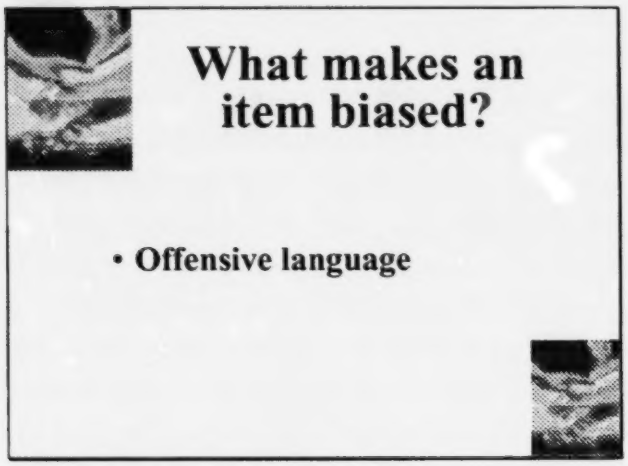
Not all real or imagined problems with test items constitute bias.

A poorly constructed item will be equally poor for everyone. Bad wording, lack of clarity, or failure to match the specifications simply make an item bad. It should be eliminated or repaired and then (if it stays in the pool) checked for bias.

Someone will take offense to just about anything. We can't eliminate a passage about timber wolves just because someone's grandmother was attacked by timber wolves as a young girl. Bias is about groups, not individuals. Things to which individual committee members object should be reevaluated in light of how groups of students might respond.

If the item addresses a specific element of instruction, it is valid. If some teachers have failed to teach that part of the curriculum, the item is not biased. The item review should serve as a heads-up to everyone to remind colleagues that the entire set of published standards will be tested. What would constitute bias would be a math item, for example, that assessed a geometric concept but could only be solved algebraically. If students who have completed all math requirements without taking advanced algebra could not be expected to answer the item correctly, it is biased against general math students.

The fact that one group of students scores higher than another group does not make the item biased. To prove bias, we would have to show that group membership overrides ability in determining how one performs on the item.



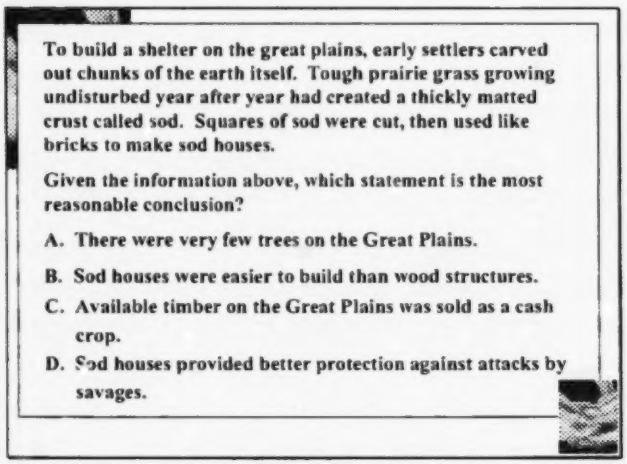
The key to all four of these sources of bias is group. We are looking for things that would bias test results for or against a group of students. There will be individual students who will be offended by just about anything, will have access to very little information, and have few interests. Everything is going to be difficult for them. We're not talking about them.

Offensive language is any word or phrase that is generally regarded as offensive to a group. Ethnic slurs, gender slurs, racial slurs, and the like would qualify, as would language that derided any group: racial, ethnic, gender, religious, regional, urban/rural, socioeconomic. Example: Hillbilly, redneck, city slicker, country bumpkin. Nonparallel construction can also be construed as offensive: men and girls, pilots and stewardesses,

Stereotype involves reinforcement of stereotypical roles: females involved only in domestic activities, minorities engaged only in menial jobs, only white males portrayed in executive or decision-making decisions. Note that white males can occasionally hold executive positions or make decisions. The issue is balance.

Differential access: If a passage or item requires special knowledge that one group is more likely to have access to, that passage or item is biased. Example, a reading passage describes lawnmower maintenance. An item that asks the student to draw a conclusion about failure to keep the air filter clean would be biased if the passage didn't contain sufficient information for a previously uninformed person to answer the item correctly. If boys are more likely to have experience maintaining lawnmowers, the item would be biased in favor of boys or against girls.

Differential interest: This is more subtle than differential access in that it is subject to self selection, which brings biases of the test takers to bear. Students tend to try harder on things in which they are interested. Therefore, if the items are skewed in a particular direction, those students with greater interest in those topics are likely to perform better on the test. This is an item bias issue but more a test balance issue.



To build a shelter on the great plains, early settlers carved out chunks of the earth itself. Tough prairie grass growing undisturbed year after year had created a thickly matted crust called sod. Squares of sod were cut, then used like bricks to make sod houses.

Given the information above, which statement is the most reasonable conclusion?

A. There were very few trees on the Great Plains.

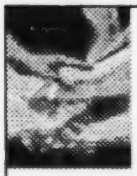
B. Sod houses were easier to build than wood structures.

C. Available timber on the Great Plains was sold as a cash crop.

D. Sod houses provided better protection against attacks by savages.

This item seems innocent enough…until we get to the very last word in option D. Since D is not the correct answer, it would seem that this would not matter. It does matter for two reasons:

1. Not everyone will know that D is not the correct answer
2. Whether D is correct or not, the test introduces the notion that non-white plains inhabitants (i.e., Native Americans) are savages, a notion that would no doubt offend many Native Americans and others as well. D should be replaced.



What makes an item biased?

- **Stereotypes**

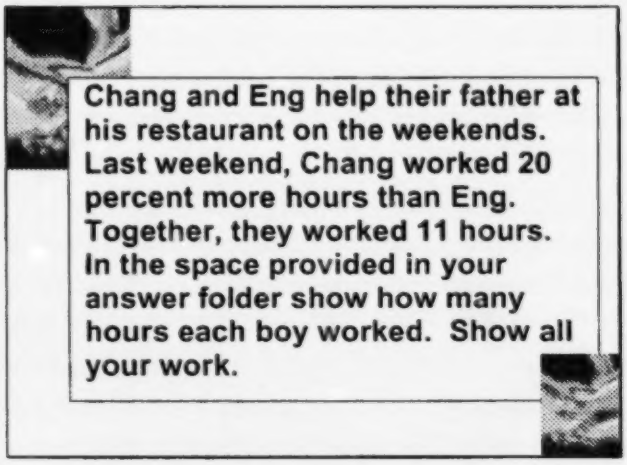
The key to all four of these sources of bias is group. We are looking for things that would bias test results for or against a group of students. There will be individual students who will be offended by just about anything, will have access to very little information, and have few interests. Everything is going to be difficult for them. We're not talking about them.

Offensive language is any word or phrase that is generally regarded as offensive to a group. Ethnic slurs, gender slurs, racial slurs, and the like would qualify, as would language that derided any group: racial, ethnic, gender, religious, regional, urban/rural, socioeconomic. Example: Hillbilly, redneck, city slicker, country bumpkin. Nonparallel construction can also be construed as offensive: men and girls, pilots and stewardesses,

Stereotype involves reinforcement of stereotypical roles: females involved only in domestic activities, minorities engaged only in menial jobs, only white males portrayed in executive or decision-making decisions. Note that white males can occasionally hold executive positions or make decisions. The issue is balance.

Differential access: If a passage or item requires special knowledge that one group is more likely to have access to, that passage or item is biased. Example, a reading passage describes lawnmower maintenance. An item that asks the student to draw a conclusion about failure to keep the air filter clean would be biased if the passage didn't contain sufficient information for a previously uninformed person to answer the item correctly. If boys are more likely to have experience maintaining lawnmowers, the item would be biased in favor of boys or against girls.

Differential interest: This is more subtle than differential access in that it is subject to self selection, which brings biases of the test takers to bear. Students tend to try harder on things in which they are interested. Therefore, if the items are skewed in a particular direction, those students with greater interest in those topics are likely to perform better on the test. This is an item bias issue but more a test balance issue.


Chang and Eng help their father at his restaurant on the weekends. Last weekend, Chang worked 20 percent more hours than Eng. Together, they worked 11 hours. In the space provided in your answer folder show how many hours each boy worked. Show all your work.

Asian Americans are permitted to do things other than operate restaurants and laundries. There are so many other ways to address this mathematical concept that it seems ludicrous to waste it on such an arch-stereotype.

What makes an item biased?

- **"Private" knowledge**

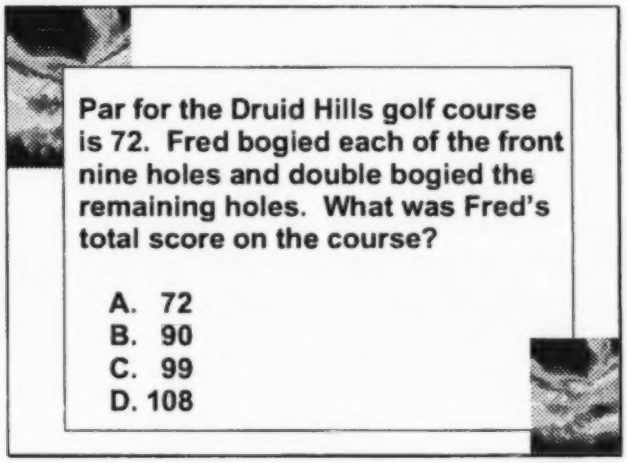
The key to all four of these sources of bias is group. We are looking for things that would bias test results for or against a group of students. There will be individual students who will be offended by just about anything, will have access to very little information, and have few interests. Everything is going to be difficult for them. We're not talking about them.

Offensive language is any word or phrase that is generally regarded as offensive to a group. Ethnic slurs, gender slurs, racial slurs, and the like would qualify, as would language that derided any group: racial, ethnic, gender, religious, regional, urban/rural, socioeconomic. Example: Hillbilly, redneck, city slicker, country bumpkin. Nonparallel construction can also be construed as offensive: men and girls, pilots and stewardesses,

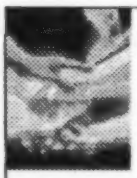
Stereotype involves reinforcement of stereotypical roles: females involved only in domestic activities, minorities engaged only in menial jobs, only white males portrayed in executive or decision-making decisions. Note that white males can occasionally hold executive positions or make decisions. The issue is balance.

Differential access: If a passage or item requires special knowledge that one group is more likely to have access to, that passage or item is biased. Example, a reading passage describes lawnmower maintenance. An item that asks the student to draw a conclusion about failure to keep the air filter clean would be biased if the passage didn't contain sufficient information for a previously uninformed person to answer the item correctly. If boys are more likely to have experience maintaining lawnmowers, the item would be biased in favor of boys or against girls.

Differential interest: This is more subtle than differential access in that it is subject to self selection, which brings biases of the test takers to bear. Students tend to try harder on things in which they are interested. Therefore, if the items are skewed in a particular direction, those students with greater interest in those topics are likely to perform better on the test. This is an item bias issue but more a test balance issue.


Par for the Druid Hills golf course is 72. Fred bogied each of the front nine holes and double bogied the remaining holes. What was Fred's total score on the course?
A. 72
B. 90
C. 99
D. 108

We could have used bridge or bowling or any other activity in which a score is kept and in which scorekeeping is something of a secret ritual. The private knowledge here, of course, includes par, bogie, and even knowing how many holes there are in regulation play. Score one for the country club set and zero for everyone else.



What are the different types of item bias?

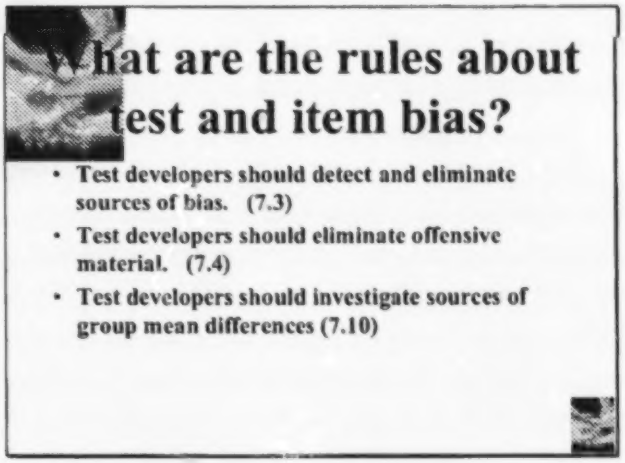
- Gender
- Racial/Ethnic
- Religious
- Age
- Socioeconomic
- Cultural
- Disability

Regional - refers to regions of the country, not likely to be a problem in a statewide test but of major interest in a national test. In large states, however, the coastal vs. mountain or upstate vs. downstate issue could arise.

Urban/rural - a variation on the theme of differential access. Statewide tests should generally not have items about subways, mass transit, irrigation, grain futures, and the like. Occasionally these topics can turn up in reading passages if all the necessary information is in the passage.

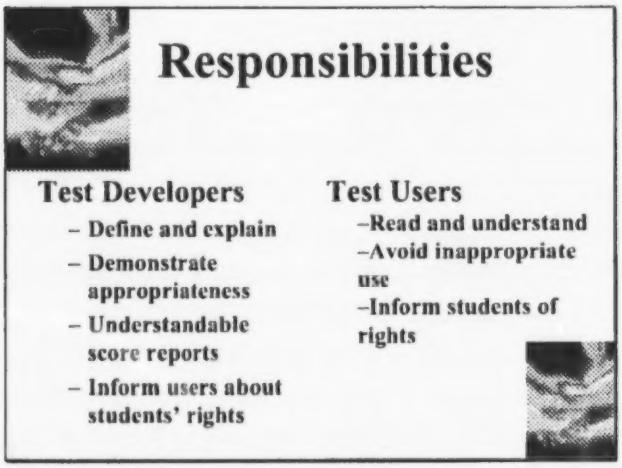
Socioeconomic - We don't want to pit the rich against the poor. This includes topics such as calculator access to items about home computers, video games, swimming pools, vacations, and other 'luxury' items and activities. In Rhode Island several years ago, we had to take a sailboat off the cover of a test because the Dept. thought some people would consider it elitist.

Religious - Evolution, human reproduction, dietary restrictions, and similar topics fall under this umbrella.

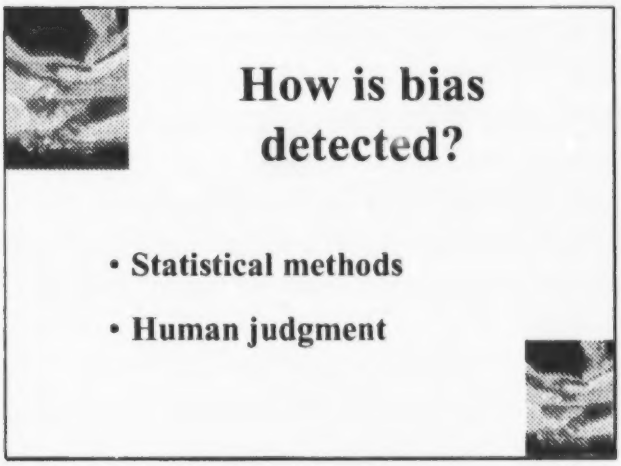

What are the rules about test and item bias?
• Test developers should detect and eliminate sources of bias. (7.3)
• Test developers should eliminate offensive material. (7.4)
• Test developers should investigate sources of group mean differences (7.10)

These are the applicable standards from the 1999 edition of *Standards for Eductional and Psychological Testing* published by AERA, APA, and NCME. Make sure to show a copy of the book and remind committee members that everyone who reviews items at MI has a copy of this book on his or her desk and that we encourage all our staff to become very familiar with it.

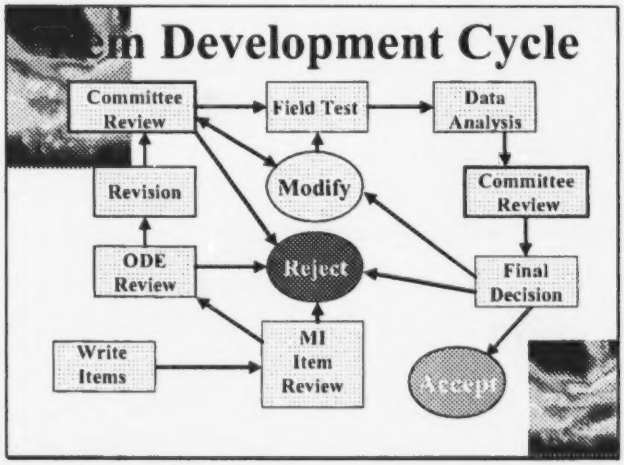
This discussion should reinforce the fact that we give every item a thorough going over before they see it. In essence, they review the items to see how well we've done our job and to add the perspective of the Ohio classroom teacher who knows Ohio students.



These responsibilities are taken directly from the Code of Fair Testing Practices. You have those in your packets. As members of this committee, it is your responsibility to assist the Department in making sure tests are fair. The Code contains a wealth of information to help you do your job. You should take some time and become as familiar with it as possible.



We use statistical methods to flag items that might be biased but human judgment to make a final determination. On the first pass, we will not have any data; therefore, human judgment is all we will have to work with.



From start to finish, it takes about two years to bring a single test item to maturity. This slide, and the next two, provide an overview and some details of that two-year journey.

MI staff review items thoroughly prior to the first of two committee meetings. After committees review the items, we make all the necessary modifications or eliminate those items that just don't work. Then we send them along to ODE for review. Once ODE approves the items for field test, we arrange them into short embedded forms to administer along with the spring test. We collect the data, analyze it, and present our findings, along with the items, to the committees a second time, about a year after their first review.

At the second review, we have a final chance to evaluate each item, this time with student performance data. Three decisions are available: Accept, Modify, Reject. If we accept an item, it goes into a pool and is eligible for inclusion in a future operational test. If we elect to modify an item, it goes back into the cycle. A rejected item will never be seen or heard from again.

The initial review at MI is extensive:

The **content specialist** reviews the item for both content and bias.

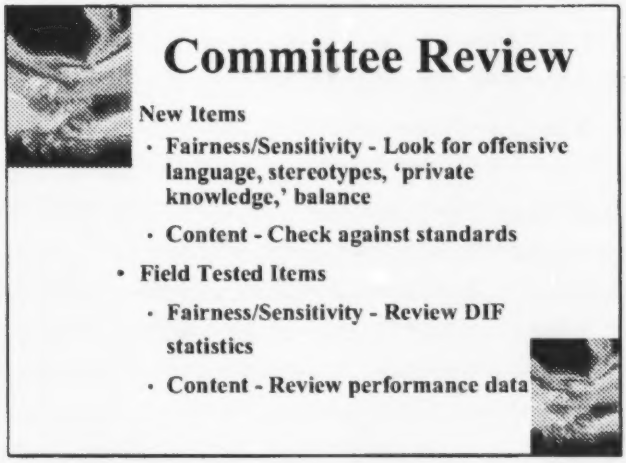
The **editor** checks for adherence to a style guide and also checks for offensive wording, stereotyping, and lack of balance.

The **project director** checks the content specialist's and editor's work.

As necessary, **senior managers** at MI check behind the project manager.

We also involve **ODE** and **university consultants**, just to make sure.

Sometimes an item goes through several rounds of revision and review within MI.


Committee Review
New Items
• Fairness/Sensitivity - Look for offensive language, stereotypes, 'private knowledge,' balance
• Content - Check against standards
• Field Tested Items
• Fairness/Sensitivity - Review DIF statistics
• Content - Review performance data

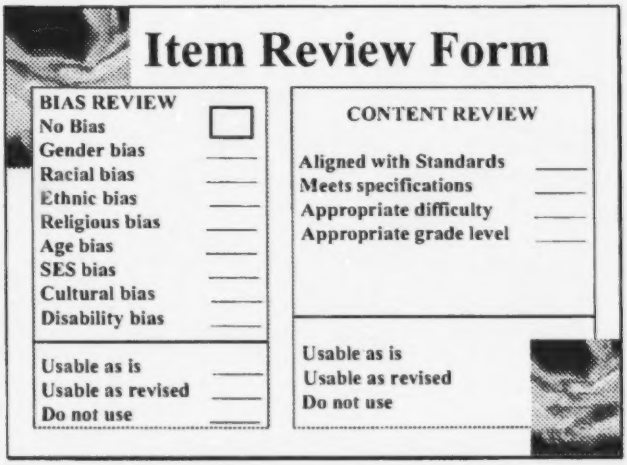
The committees get two chances to review each item, once without data, and once with:

At the first review, we will examine and discuss each item in terms of its adherence to the specifications and presence of offensive wording, stereotypes, private knowledge, and differential interest.

At the second review, we present the final version of the item, as it was field tested, along with data:

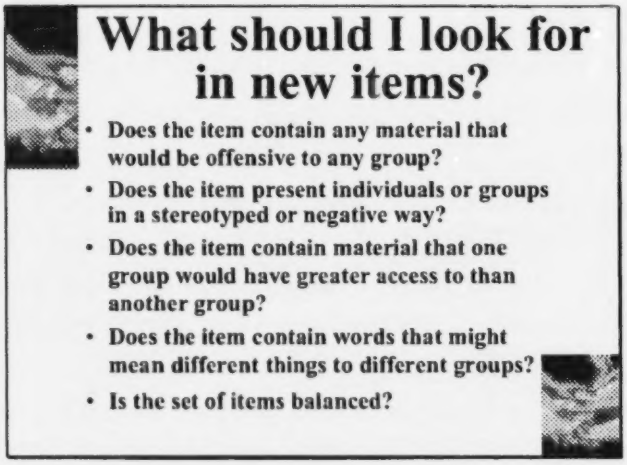
a. P value
b. Response frequency distribution
c. Point biserial correlation with total score
d. a-d by sex and race
e. Mantel-Haenszel DIF statistics (explained in the next slide)

Whether it is the first or second review, bias is determined solely by committee members; the statistics and our comments are to be considered resource material only.



This is an example of the review form you will use to evaluate individual test items. Note the large box at the top of the Bias Review side of the form. If the item has no bias issues, check this box. If you cannot check this box, indicate below which issues (one or more) keep this item from being bias-free.

You will also notice that the content review committees will use the same form. They are responsible for determining whether or not the item adequately addresses the Standards and Benchmarks at an appropriate level of difficulty. You may have comments or concerns about content and difficulty. Feel free to raise them, but be assured that the content committees will be responsible for those issues. Similarly, members of the content review committees may have concerns about potential bias. They will raise those issues, but you will have responsibility for doing something about them.


What should I look for in new items?
• Does the item contain any material that would be offensive to any group?
• Does the item present individuals or groups in a stereotyped or negative way?
• Does the item contain material that one group would have greater access to than another group?
• Does the item contain words that might mean different things to different groups?
• Is the set of items balanced?

Balance refers to a complete set of items. Such a set should be balanced with respect to names, roles, and interests. If you review a set of 50 math items, and it seems that boys are constantly outdoors doing active things and girls are always indoors doing passive things, we need to make some changes in the items.

What should I look for in field tested items?

- Are there group differences on the base test?
- Did all groups perform about the same on this item?
- Did one distractor attract one group more than others?
- Is there significant differential item functioning?

These are reviews of data. MI will summarize the data and provide a thorough explanation of what each statistic means. We will also show committee members how to use the data and how to track from one statistic to another in search of meaning.

What is differential item functioning (DIF)?

A systematic difference between two groups in performance on a given test item that cannot be explained on the basis of known differences in the abilities of the two groups

This is a statistical term. It is not the same as bias.

Is DIF the same as bias?

- No
- DIF is a statistical result.
- Bias can only be detected by human judgment.

This should be self explanatory.

How is differential item functioning measured?

- **Mantel-Haenszel Chi Square**
- **Mantel-Haenszel Categories**
  - **A No DIF**
  - **B Possible DIF**
  - **C Probable DIF**

We will do the work, give you the results, and let you make the decisions. The results will break down items into three groups:

A No DIF
B Possible DIF
C Definite DIF

While all the C items should be reviewed, there is no constraint against reviewing B or A items.

At a given ability level, operationally defined here as score on the base test, we would expect two groups of students to perform about the same on any new item. For example, if we took all the white students who got 50 items right on the base test and all the black students who got 50 items right on the base test, we would expect those two groups of students to have about the same level of ability. Therefore, we would expect comparable percentages of both groups to answer item 51 (a field test item) correctly. If 74% of white students answered item 51 correctly and 63% of black students answered item 51 correctly, we would note a difference of 11%. We would then look at all the students who got scores of 49 on the base test, compare the performances of black and white students on item 51, and enter the difference. We would repeat this comparison for every base test raw score group. We would then tally the absolute value of the differences. It is the size of this summed difference that determines DIF. We can also look at DIF just at the high end of ability or low end or middle.

Procedures

- **Sign non-disclosure forms**
- **Receive training from MI staff**
- **Receive items/review forms**
- **Review items alone**
- **Review items as a group**
- **Sign off**

Your task as members of the Fairness/Sensitivity Committee will be to review new and field tested items, using information and forms we supply. Each time you meet, you will do several things:

First, you will sign a non-disclosure form, stating that you will observe and abide by all test security rules. Then, depending on the exact task you will perform, we will provide an orientation and specific training.
Next you will receive test items, typically organized in a way that facilitates orderly review. You will receive the review form at the same time. You will initially review a set of items alone, and then the group will discuss the items, either one at a time or in an order agreed upon by the Committee.
Finally, after discussing the items, the Committee chair will poll the members of the Committee to determine that all criteria are met or that specific criteria are not met and sign the form for each item. Should any item fail to meet any of the criteria, the chair will record the consensus recommendation of the Committee.

To receive an electronic copy of this appendix via e-mail, please contact adechant@measinc.com . Specify Powerpoint (1.7 megabytes) or Adobe Acrobat (1.6 megabytes). Make sure your e-mail server supports transfer of files this size.

BEST COPY AVAILABLE

TM034657

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC

Reproduction Release

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Item Review 101: Where We've Been, Where We're Going, How We'll Get There

Author(s): Michael B. Bunch

Corporate Source: Measurement Incorporated

Publication Date: June 2002

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, Resources in Education (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign in the indicated space following.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
Level 1	**Level 2A**	**Level 2B**
☑	☐	☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g. electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche, or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature: Michael B. Bunch

Printed Name/Position/Title: Michael B. Bunch, Ph.D., Vice President

Organization/Address:
Measurement Incorporated
423 Morris Street
Durham, NC 27701

Telephone: (919) 683-2413

Fax: (775) 257-2328

E-mail Address: mbunch@measinc.com

Date: September 10, 2002

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:	
ERIC Clearinghouse on Assessment and Evaluation **1129 Shriver Laboratory (Bldg 075)** **College Park, Maryland 20742**	**Telephone: 301-405-7449** **Toll Free: 800-464-3742** **Fax: 301-405-8134** **ericae@ericae.net** **http://ericae.net**

EFF-088 (Rev. 9/97)